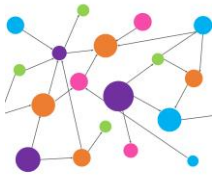
[CATI ~]$ grep

CATI

Groupe de Travail Graphes
de connaissances



Séminaire IN-OVIVE Linked Data

Intégration de données hétérogènes dans une base de données graphe

Flores R, Confais J, Francillonne N, Toumert Y, Destin J, Bogoin J, Michotey C, Pommier C, Alfama F, Quesneville H, Alaux M (URGI), Rimbert H (GDEC), Tamby JP (IPS2), Kreplak J, Imbert B (UMR Agroécologie) – CATI GREP

Sète, 2021-10-12

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

Historique

Collaboration GDEC-URGI SyntenyViewer (ANR 2011 puis 2011-2017)

- SGBDR données de synténie, dataset Cereals
- Gènes, positions chromosomiques, homologies, taxon

Plant2Pro SyntenyViewer (2017-2021)

- SGBDR données de synténie, familles botaniques
- POC LPG (Labeled Property Graph) avec Neo4J

CATI GREP (2018-en cours)

- Groupe de travail novice en technologies graphe et sémantique
- Étendre l'usage de Neo4J pour une intégration de données hétérogènes plus large

RDF vs Graphe de propriété ?

RDF/Turtle

RDF vs Graphe de propriété ?

RDF/Turtle

```
@base <http://example.org> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
```

RDF vs Graphe de propriété ?

RDF/Turtle

```
@base <http://example.org> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
<#Alice> a foaf:Person;
```

RDF vs Graphe de propriété ?

RDF/Turtle

```
@base <http://example.org> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
<#Alice> a foaf:Person;
    foaf:account_name @alice;
    foaf:name "Alice Smith";
    foaf:age 42.
```

RDF vs Graphe de propriété ?

RDF/Turtle

```
@base <http://example.org> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
<#Alice> a foaf:Person;
    foaf:account_name @alice;
    foaf:name "Alice Smith";
    foaf:age 42.
<#Bob> a foaf:Person;
```

RDF vs Graphe de propriété ?

RDF/Turtle

```
@base <http://example.org> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
<#Alice> a foaf:Person;
    foaf:account_name @alice;
    foaf:name "Alice Smith";
    foaf:age 42.
<#Bob> a foaf:Person;
    foaf:knows <#Alice>.
```

RDF vs Graphe de propriété ?

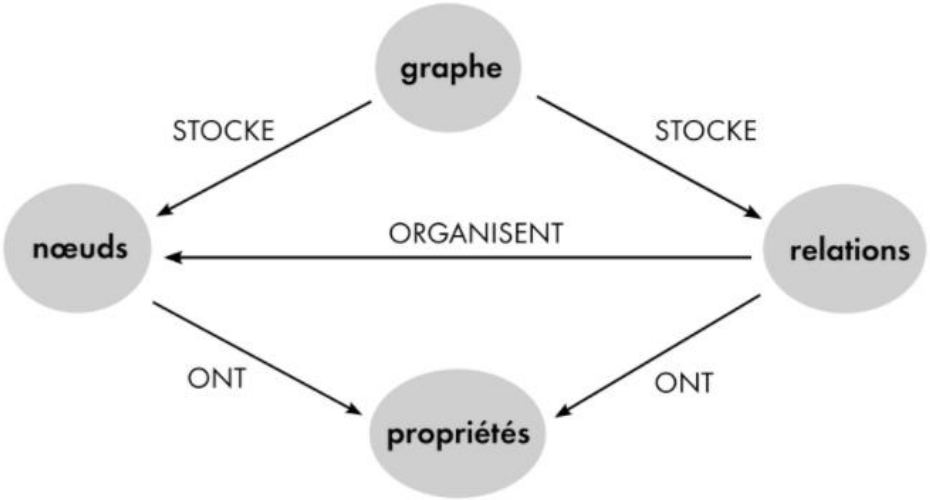
RDF/Turtle

```
@base <http://example.org> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
<#Alice> a foaf:Person;
    foaf:account_name @alice;
    foaf:name "Alice Smith";
    foaf:age 42.
<#Bob> a foaf:Person;
    foaf:knows <#Alice>.
```

Pas de structure interne aux nœuds ou relations

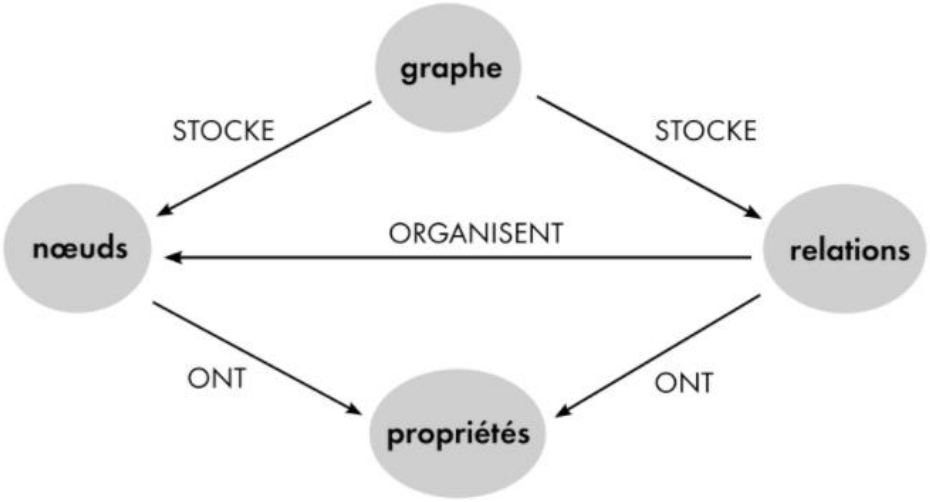
RDF vs Graphe de propriété ?

Labeled Property Graph - LPG



RDF vs Graphe de propriété ?

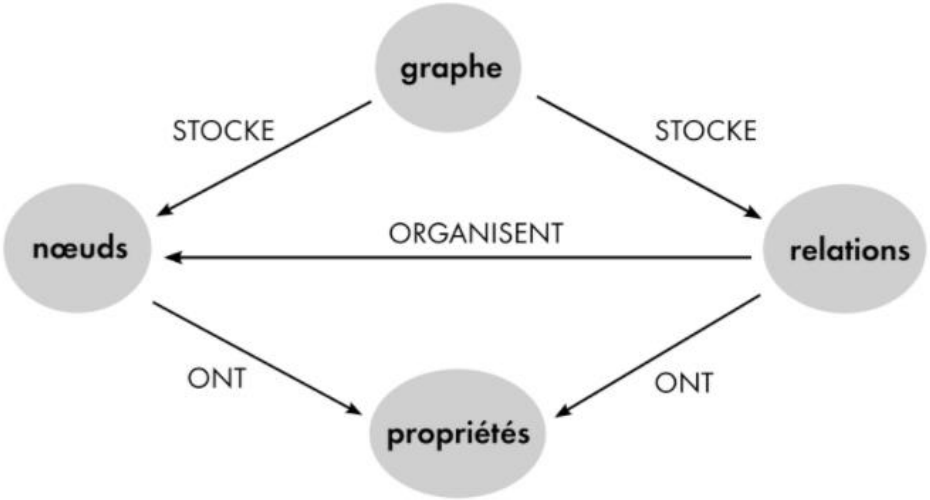
Labeled Property Graph - LPG




```
(node { property: “value”} )-[:RELATION {property: “value”}]->(node)
```

RDF vs Graphe de propriété ?

Labeled Property Graph - LPG




```
(Bob:Person)
-[:KNOWS]->
(Alice:Person { account_name: "@alice", name: "Alice Smith", age: 42 })
```

```
(node { property: “value”} )-[:RELATION {property: “value”}]->(node)
```

RDF vs Graphe de propriété ?

Quelques différences

- RDF :
  - modèle généralement additif
  - CRUD, mais update = delete/insert
- Neo4J :
  - Modèle dynamique - CRUD + DML (Data Manipulation Language)
  - Méta-modèle généré à la volée à partir des données
  - Cypher simplifie la prise en main
  - Stockage graphe natif : relations directement matérialisées, non calculées lors de la requête => performances accrues

Forces de Neo4J

Écosystème permettant de manipuler les données avec fluidité

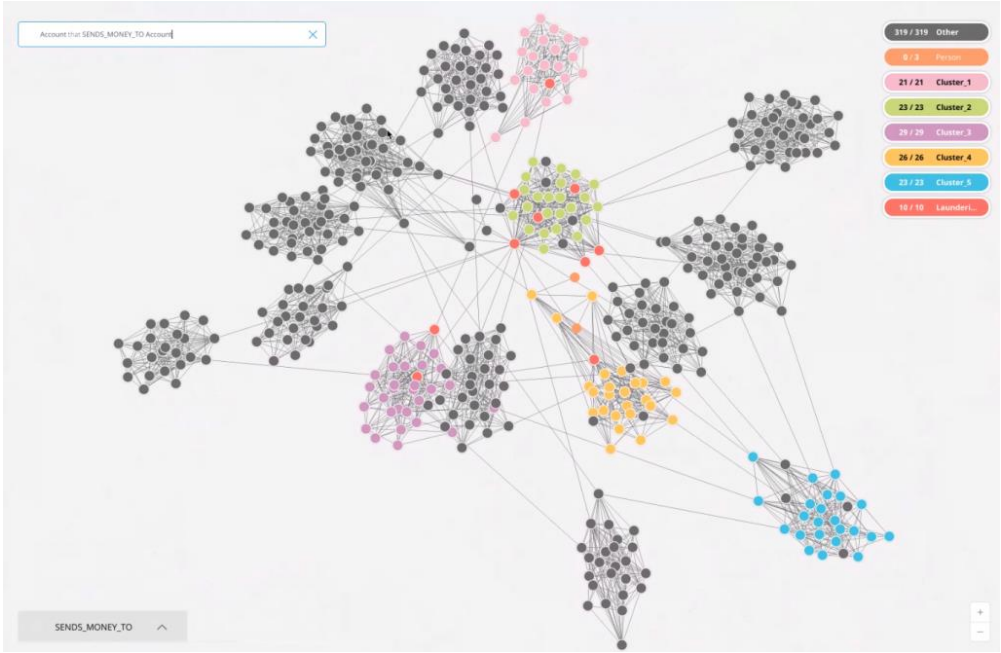
- Navigation et exploration/visualisation
  - Neo4J browser
  - Bloom

- Bibliothèque APOC
  - Import/export de données
  - Procédures utilitaires (conversion, calculs...)
  - Refactoring de graphes

Forces de Neo4J

Écosystème permettant de manipuler les données avec fluidité

- Navigation et exploration/visualisation
  - Neo4J browser
  - Bloom



- Bibliothèque APOC
  - Import/export de données
  - Procédures utilitaires (conversion, calculs...)
  - Refactoring de graphes

Forces de Neo4J

Graph Data Science / Algorithmes

Forces de Neo4J

Graph Data Science / Algorithmes

– Parcours de graphe : trouver le chemin le plus court entre deux concepts ou évaluer la disponibilité ou la qualité d'un chemin.

Forces de Neo4J

Graph Data Science / Algorithmes

– Parcours de graphe : trouver le chemin le plus court entre deux concepts ou évaluer la disponibilité ou la qualité d'un chemin.

=> existe-t-il un lien fort entre 2 phénotypes d'intérêt impliquant un unique gène ? Permettrait de guider un travail de sélection variétale ?

Forces de Neo4J

Graph Data Science / Algorithmes

– Parcours de graphe : trouver le chemin le plus court entre deux concepts ou évaluer la disponibilité ou la qualité d'un chemin.

=> existe-t-il un lien fort entre 2 phénotypes d'intérêt impliquant un unique gène ? Permettrait de guider un travail de sélection variétale ?

– Centralité : déterminer l'importance de nœuds dans le graphe.

Forces de Neo4J

Graph Data Science / Algorithmes

– Parcours de graphe : trouver le chemin le plus court entre deux concepts ou évaluer la disponibilité ou la qualité d'un chemin.

=> existe-t-il un lien fort entre 2 phénotypes d'intérêt impliquant un unique gène ? Permettrait de guider un travail de sélection variétale ?

– Centralité : déterminer l'importance de nœuds dans le graphe.

=> identifier quels gènes sont le plus impliqués dans la réponse à un stress ou à un processus

Forces de Neo4J

Graph Data Science / Algorithmes

– Parcours de graphe : trouver le chemin le plus court entre deux concepts ou évaluer la disponibilité ou la qualité d'un chemin.

=> existe-t-il un lien fort entre 2 phénotypes d'intérêt impliquant un unique gène ? Permettrait de guider un travail de sélection variétale ?

– Centralité : déterminer l'importance de nœuds dans le graphe.

=> identifier quels gènes sont le plus impliqués dans la réponse à un stress ou à un processus

– Détection de communautés : évaluer à quel point un groupe est partitionné.

Forces de Neo4J

Graph Data Science / Algorithmes

– Parcours de graphe : trouver le chemin le plus court entre deux concepts ou évaluer la disponibilité ou la qualité d'un chemin.

=> existe-t-il un lien fort entre 2 phénotypes d'intérêt impliquant un unique gène ? Permettrait de guider un travail de sélection variétale ?

– Centralité : déterminer l'importance de nœuds dans le graphe.

=> identifier quels gènes sont le plus impliqués dans la réponse à un stress ou à un processus

– Détection de communautés : évaluer à quel point un groupe est partitionné.

=> identifier quelles protéines ou métabolites intéragissent ensembles dans une réponse à un stress ou un processus biologique

Inconvénients de Neo4J

- Édition communautaire manque de certaines fonctionnalités

  - Import performant possible uniquement si base vide

  - Exploitation matérielle bridée à 4 processeurs pour certaines fonctions

  - Scalabilité (extensibilité ?) horizontale non gérée

Inconvénients de Neo4J

- Édition communautaire manque de certaines fonctionnalités

  - Import performant possible uniquement si base vide

  - Exploitation matérielle bridée à 4 processeurs pour certaines fonctions

  - Scalabilité (extensibilité ?) horizontale non gérée

- Coût licence Entreprise assez exorbitant (aux dernières nouvelles lointaines)

  cible pas mal le monde de la finance et l'industrie.

  https://neo4j.com/pricing <= peu d'info, y'a un loup

Processus d'intégration

Données éparses, formatage non graphe

Traitements et développements de scripts *ad hoc*.

Il faut identifier les questions pour modéliser correctement les données.

1. Approche externe
    - Génération de CSV correspondant à l'ensemble des nœuds, relations, propriétés
    - Règles gérées en dehors de la base
    - Import performant via Neo4J admin (uniquement à partir d'une base vide)
2. Approche hybride
    - Chargement brut des données dans Neo4J
    - Exploitation des fonctionnalités de refactoring de graphe pour construire les relations à partir des propriétés

Données RDF ou graphe

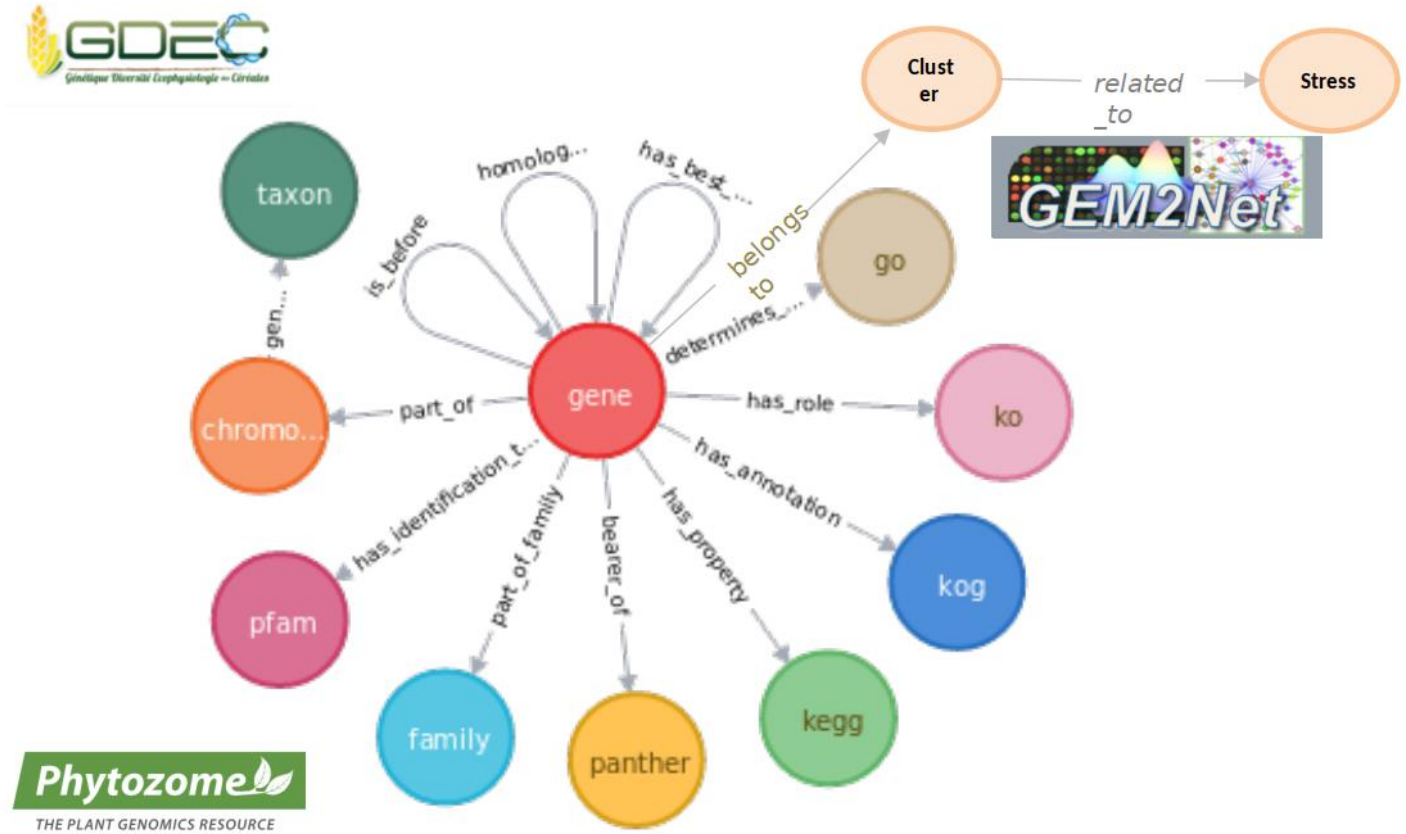
– Import direct dans Neo4J via dump/restore, ou plugin Neosemantics (n10s)

– Refactoring de graphe pour exploiter les connaissances disponibles et les lier entre elles si besoin

Cas d'utilisation

Données d'homologie niveau angiospermes & profils d'expression

Annotations GO Slim / Lien Arabidopsis thaliana



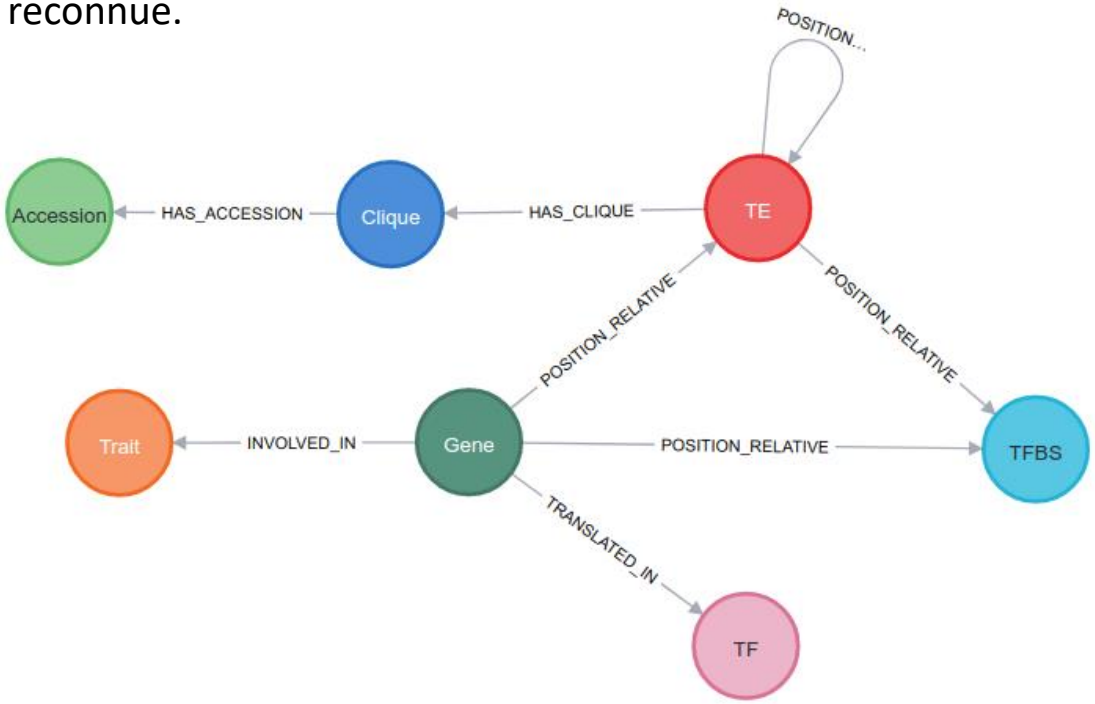
Cas d'utilisation

Étude des pangénomes de TE d'Arabidopsis thaliana (espèce modèle de plantes)

Étude co-localisation entre entités génomiques (BedTools), impact des TE (éléments transposables) sur les phénotypes exprimés dans certains variétés/accessions.

Difficulté de trouver des termes d'ontologie décrivant sans ambiguïté les relations entre objets.

Utilisation de termes *ad hoc* pour décrire les relations, sans prétention d'en faire une ontologie reconnue.



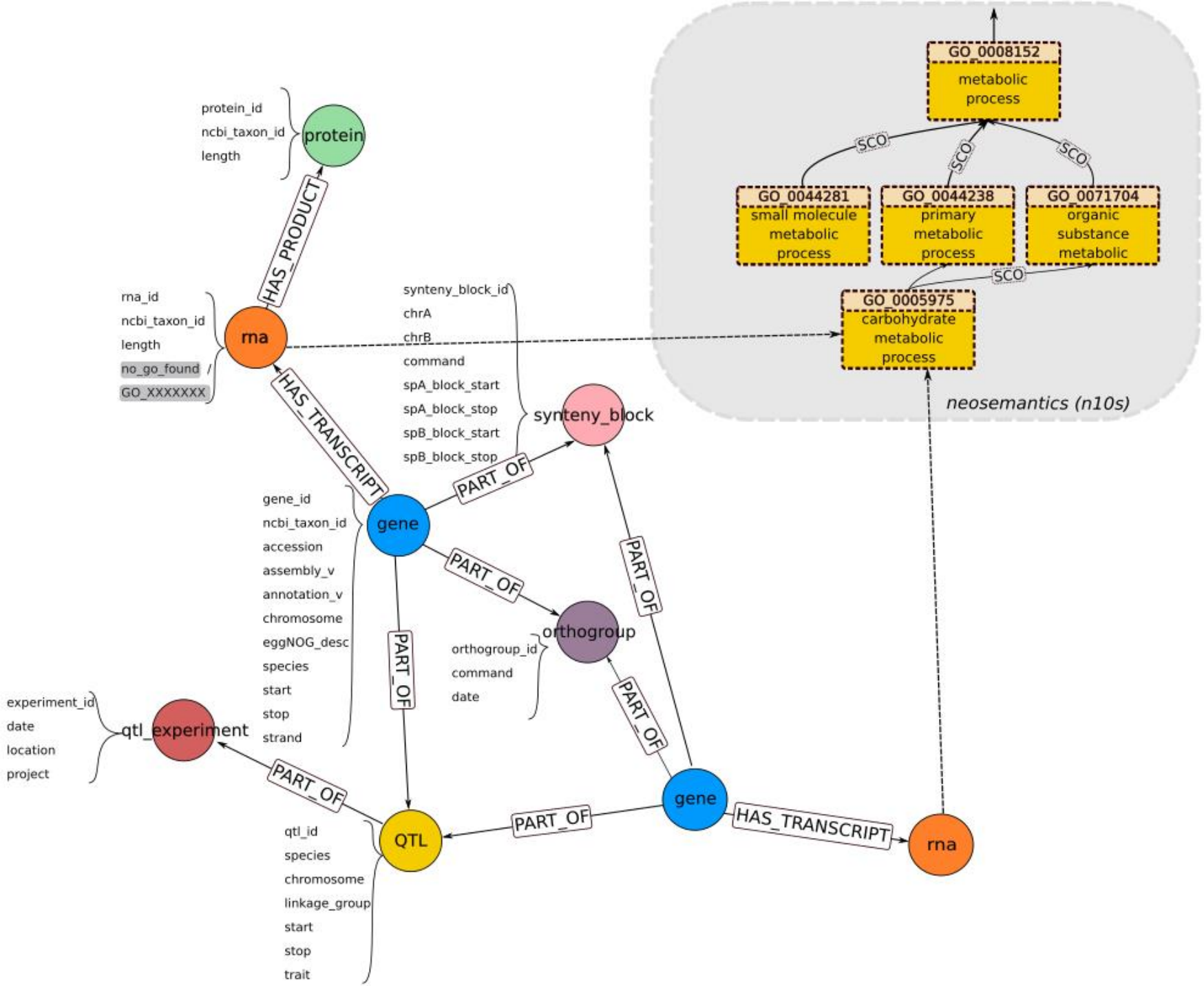
Cas d'utilisation

Recherche translationnelle chez les légumineuses

- Construction de groupes d'orthologues légumineuses spécifiques via OrthoFinder
  - Orthologues : gènes homologues ayant subi un événement de spéciation
  - les gènes orthologues peuvent conservent leur fonction au cours de l'évolution

    => inférence de fonctions d'un gène d'une espèce vers celui d'une autre espèce
- Identification de blocs de synténie (conservation ordre des gènes) entre espèces via DAGchainer
  - 4 gènes consécutifs
  - Distance de moins d'1 Mb entre chaque gène
- Intégration de données génétiques
  - QTL
  - GWAS
- Intégration avec la Gene Ontology (via neosemantics)

  => inférence de connaissances


protein_id
ncbi_taxon_id
length
protein
HAS_PRODUCT
rna_id
ncbi_taxon_id
length
no_go_found /
GO_XXXXXXX
rna
HAS_TRANSCRIPT
synteny_block_id
chrA
chrB
command
spA_block_start
spA_block_stop
spB_block_start
spB_block_stop
synteny_block
PART_OF
gene_id
ncbi_taxon_id
accession
assembly_v
annotation_v
chromosome
eggNOG_desc
species
start
stop
strand
gene
PART_OF
PART_OF
orthogroup_id
command
date
orthogroup
PART_OF
PART_OF
experiment_id
date
location
project
qtl_experiment
PART_OF
qtl_id
species
chromosome
linkage_group
start
stop
trait
QTL
PART_OF
gene
HAS_TRANSCRIPT
rna
GO_0008152
metabolic process
SCO
SCO
SCO
GO_0044281
small molecule metabolic process
GO_0044238
primary metabolic process
GO_0071704
organic substance metabolic
SCO
GO_0005975
carbohydrate metabolic process
neosemantics (n10s)

Bilan données intégrées

- Multi-espèces (angiospermes, Arabidopsis thaliana, blé, riz, vigne, peuplier, chêne)
- Échelle multi-omique (annotations de gènes, marqueurs génomiques/génétiques, QTL, transcrits/protéines, fonctions, phénotypes/traits, etc.) + littérature (à venir) mais pas de séquences !
- Données publiques + données projets en cours (sous embargo)
- Volumétrie : modérée, sources de quelques dizaines de Go zippés avec séquences, mais intégration Neo4J
  - ~Go
  - ~ 3M nœuds
  - ~ 6M relations

Quelques difficultés rencontrées

- Ontologies utilisées :
  - De référence (Gene Ontology, Sequence Ontology)
  - Ou plus spécifiques (TO, WTO, CO_321 / Wheat, CO_357 / Woody)
- La formalisation avec des ontologies peut être gérée lors de l'export RDF en fin d'intégration via un mapping n10s
- Modélisation très liée aux entités manipulées, mais aussi et surtout aux questions envisagées
- Gestion de la provenance perfectible (sic). Pistes :
  - PAV (Provenance, Authoring and Versioning) : https://pav-ontology.github.io/pav/ (plutôt léger)
  - Provenance DCTerms : https://www.dublincore.org/specifications/dublin-core/dcmi-terms/terms/provenance/
  - PROV-O : https://www.w3.org/TR/2013/REC-prov-o-20130430/ (overkill ?)

Questions en suspens...

- Quelle stratégie adopter pour l'utilisation de termes d'ontologies lors de la diffusion au format RDF ?
  - *Cherry-picking* ?
  - Extension d'ontologies existantes ?
  - Nommage *ad hoc* sans formalisme ?
  - Alimentation de termes spécifiques dans le thésaurus INRAE ? (ping Sophie *et al.*)
- Provenance des données, comment la gérer de manière fiable et pérenne ? Quelle granularité ?
  - Propriétés ?
  - Nœud ?
  - Relation ?
  - Type de nœud ou de relation ?
  - Sous-graphe complet ?
- Workshop Work4Graph'Int les 9-10/12/2021 à Paris, financement DipSO/DSI
  - Intégration de données hétérogènes en base graphe
  - Inscriptions et soumissions ouvertes :

  https://work4graph.pages.mia.inra.fr/work4graph-integration/